CLAIMS

What is claimed is:

1. A method for load balancing requests on a network, the method comprising:

receiving a request from a requestor having a requestor network address at a first load balancer having a first load balancer network address, said request having a source address indicating said requestor network address and a destination address indicating said first load balancer network address;

forwarding said request from said first load balancer to a second load balancer at a triangulation network address, said request source address indicating said requestor network address and said destination address indicating said triangulation network address, said triangulation network address being associated with said first load balancer network address; and

sending a response from said second load balancer to said requestor at said requestor network address, said response having a source address indicating said first load balancer network address associated with said triangulation network address and a destination address indicating said first requestor network address.

2. A method according to claim 1 and further comprising:

maintaining said association between said triangulation network address and said first load balancer network address at either of said load balancers.

3. A method according to claim 1 and further comprising:

maintaining said association between said triangulation network address and said first load balancer network address at said second load balancer; and

communicating said association to said first load balancer.

4. A method according to claim 1 and further comprising:

directing said request from said second load balancer to a server in communication with said second load balancer;

composing said response at said server; and

providing said response to said second load balancer.

5. A method for load balancing requests on a network, the method comprising:

determining the network proximity of a requestor with respect to each of at least two load balancers;

designating a closest one of said load balancers by ranking said load balancers by network proximity; and

directing requests from said requestor to said closest load balancer.

6. A method according to claim 5 and further comprising directing requests from any source having a subnet that is the same as the subnet of said requestor to said closest load balancer.

7. A method according to claim 5 and further comprising:

monitoring the current load of each of said load balancers; and

performing said directing step wherein the current load of said closest load balancer is less than the current load of every other of said load balancers.

8. A method according to claim 5 wherein said determining step comprises periodically determining.

9. A method according to claim 5 wherein said determining step comprises determining at at least one fixed time.

10. A method according to claim 5 wherein said determining step comprises polling said requestor to yield at least two attributes selected from the group consisting of: latency, relative TTL, and number of hops to requestor.

11. A method according to claim 5 wherein said determining step comprises polling said requestor using at least two polling methods selected from the group consisting of: pinging, sending a TCP ACK message to said requestor's source address and port, sending a TCP ACK message to said requestor's source address and port 80, and sending a UDP request to a sufficiently high port number as to elicit an "ICMP port unreachable" reply.

12. A method according to claim 5 wherein said designating step comprises designating a closest one of said load balancers by ranking said load balancers by network proximity and either of current load and available capacity.

13. A method for determining network proximity, the method comprising:

sending from each of at least two servers a UDP request having a starting TTL value to a client at a sufficiently high port number as to elicit an "ICMP port unreachable" reply message to at least one determining one of said servers indicating said UDP request's TTL value on arrival at said client;

determining a number of hops from each of said servers to said client by subtracting said starting TTL value from said TTL value on arrival for each of said servers; and

determining which of said servers has fewer hops of said client; and

designating said server having fewer hops as being closer to said client than the other of said servers.

14. A network load [redacted]ncing system comprising:

a network;

a first load balancer connected to said network and having a first load balancer network address;

a second load balancer connected to said network and having a triangulation network address, said triangulation network address being associated with said first load balancer network address; and

a requestor connected to said network and having a requestor network address,

wherein said requestor is operative to send a request via said network to said first load balancer, said request having a source address indicating said requestor network address and a destination address indicating said first load balancer network address, wherein said first load balancer is operative to forward said request to said second load balancer at said triangulation network address, said request source address indicating said requestor network address and said destination address indicating said triangulation network address, and wherein said second load balancer is operative to send a response to said requestor at said requestor network address, said response having a source address indicating said first load balancer network address associated with said triangulation network address and a destination address indicating said first requestor network address.

15. A system according to claim 14 wherein either of said load balancers is operative to maintain a table of said association between said triangulation network address and said first load balancer network address.

16. A system according to claim 14 wherein said second load balancer is operative to maintain a table of said association between said triangulation network address and said first load balancer network address and communicate said association to said first load balancer.

17. A system according to claim 14 and further comprising a server in communication with said second load balancer, wherein said second load balancer is operative to direct said request from said second load balancer to said server, and wherein said server is operative to compose said response and provide said response to said second load balancer.

18. A network load balancing system comprising:

a network;

at least two load balancers connected to said network; and

a requestor connected to said network,

wherein each of said at least two load balancers is operative to determine the network proximity of said requestor, and wherein at least one of said load balancers is operative to designate a closest one of said load balancers by ranking said load balancers by network proximity and direct requests from either of said requestor and a subnet of said requestor to said closest load balancer.

19. A system according to claim 18 wherein said load balancers are operative to poll said requestor to yield at least two attributes selected from the group consisting of: latency, relative TTL, and number of hops to requestor.

20. A system according to claim 18 wherein said load balancers are operative to poll said requestor using at least two polling methods selected from the group consisting of: pinging, sending a TCP ACK message to said requestor's source address and port, sending a TCP ACK message to said requestor's source address and port 80, and sending a UDP request to a sufficiently high port number as to elicit an "ICMP port unreachable" reply.

21. A system according to claim 18 wherein said at least one of said load balancers is operative to designate said closest one of said load balancers by ranking said load balancers by network proximity and either of current load and available capacity.